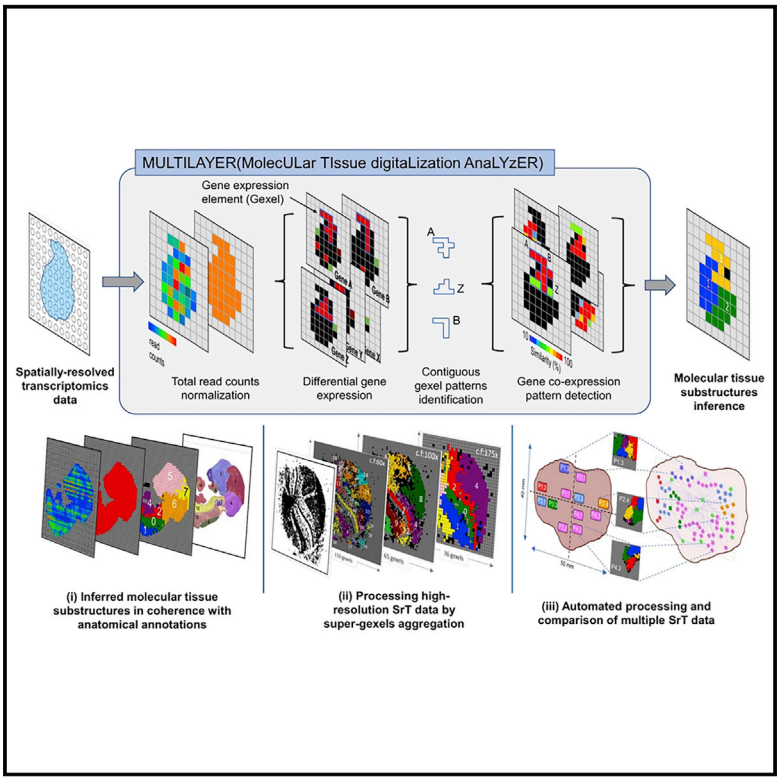
Methods

Cell Systems

# Inferring biologically relevant molecular tissue substructures by agglomerative clustering of digitized spatial transcriptomes with multilayer

## Graphical abstract




## Authors

Julien Moehlin, Bastien Mollet, Bruno Maria Colombo, Marco Antonio Mendoza-Parra

## Correspondence

mmendoza@genoscope.cns.fr

## In brief

Current methods to analyze spatially resolved transcriptomics (SrT) underexploit their spatial signature. Inspired by contextual pixel classification strategies applied to image analysis, Moehlin et al. developed MULTILAYER, a tool able to stratify SrT into functionally relevant molecular substructures. MULTILAYER proved enhanced performance on various SrT, including those of high resolution.

## Highlights

- SrT analyses underexploit their spatial signature
- MULTILAYER considers SrT as a raster image made by gexels
- MULTILAYER resolves molecular tissue substructures thanks to digital image strategies
- MULTILAYER proved enhanced performance on a variety of SrT and allows their comparison

Moehlin et al., 2021, Cell Systems *12*, 694–705
July 21, 2021 
https://doi.org/10.1016/j.cels.2021.04.008

CellPress

CellPress
OPEN ACCESS

Cell Systems

Methods

# Inferring biologically relevant molecular tissue substructures by agglomerative clustering of digitized spatial transcriptomes with multilayer

Julien Moehlin,[1] Bastien Mollet,[1,2] Bruno Maria Colombo,[1] and Marco Antonio Mendoza-Parra[1,3,*]
[1]Génomique métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057, Evry, France
[2]École Normale Supérieure de Lyon, Université Claude Bernard – Lyon 1, Université de Lyon, 69342 Lyon Cedex 07, France
[3]Lead contact
*Correspondence: mmendoza@genoscope.cns.fr
https://doi.org/10.1016/j.cels.2021.04.008

## SUMMARY

Spatially resolved transcriptomics (SrT) can investigate organ or tissue architecture from the angle of gene programs that define their molecular complexity. However, computational methods to analyze SrT data underexploit their spatial signature. Inspired by contextual pixel classification strategies applied to image analysis, we developed MULTILAYER to stratify maps into functionally relevant molecular substructures. MULTILAYER applies agglomerative clustering within contiguous locally defined transcriptomes (gene expression elements or "gexels") combined with community detection methods for graphical partitioning.
MULTILAYER resolves molecular tissue substructures within a variety of SrT data with superior performance to commonly used dimensionality reduction strategies and still detects differentially expressed genes on par with existing methods.
MULTILAYER can process high-resolution as well as multiple SrT data in a comparative mode, anticipating future needs in the field. MULTILAYER provides a digital image perspective for SrT analysis and opens the door to contextual gexel classification strategies for developing self-supervised molecular diagnosis solutions.
A record of this paper's transparent peer review process is included in the supplemental information.

## INTRODUCTION

One of the current challenges of systems biology is the study of complex living systems by evaluation of the various gene programs that define organ/tissue architecture. Indeed, accessing the gene programs in tissues has, until recently, been performed by global (bulk) gene expression analyses but recent advances in single-cell transcriptomics has made it possible to move from an "average view" toward single-cell gene-program readouts (Birnbaum, 2018). However, cell dissociation by enzymatic methods, necessary for single-cell assays, tends to modify transcriptional patterns (van den Brink et al., 2017), it destroys at least a fraction of the cells that compose the tissue and does not conserve tissue architecture.

Recent developments in spatially resolved transcriptomics (SrT) (Liu et al., 2020; Rodriques et al., 2019; Ståhl et al., 2016) have made it possible to circumvent the aforementioned technical issues related to single-cell assays, notably the capacity to conserve the spatial architecture, essential for heterogeneous tissue analysis. These strategies, based on the use of physical supports (DNA arrays [Rodriques et al., 2019; Ståhl et al., 2016] or microfluidic channels [Liu et al., 2020]) to capture local gene expression signatures (mRNA transcriptome) from tissue sections, behave like a digital camera, making it possible to obtain a "digital" view of the molecular programs of the tissue.

Although several computational solutions are available to process SrT (Bergenstråhle et al., 2020; Fernández Navarro et al., 2019), their analytical pipelines tend to reuse strategies applied to single-cell transcriptomics, namely to consider each of the captured local transcriptomes as independent units during their comparison. Specifically, the use of dimensionality reduction strategies combined with clustering methodologies have become the proven analytical path to decrease noise and facilitate data visualization within single-cell transcriptomics assays (Sun et al., 2019). This being said, recent benchmark studies demonstrated that the choice of the dimensionality reduction (e.g., PCA, t-SNE, and UMAP) and clustering methodology (e.g., K-means and hierarchical), along with their associated parameters, can give rise to divergent cell-type classifications (Becht et al., 2018; Feng et al., 2020; Raimundo et al., 2020; Sun et al., 2019), arguing for the cautious use of such methodologies (Kiselev et al., 2019). Although SrT data processed using these strategies are expected to suffer from the same pitfalls, the spatial information conserved within their captured local transcriptomes represents a major under-exploited advantage.

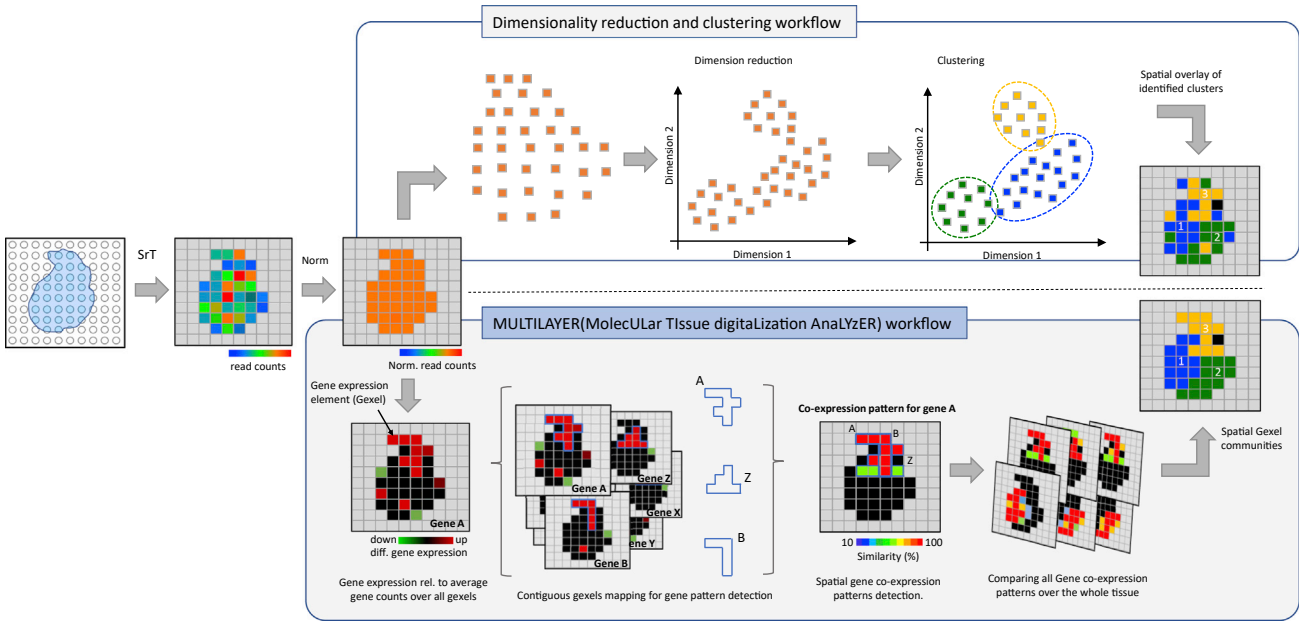



**Figure 1. The molecular tissue digitalization analyzer (MULTILAYER) workflow compared with dimensionality reduction and clustering strategies**

Spatially resolved transcriptomics (SrT) provide matrices composed of spatial coordinates harboring read counts per gene. Such coordinates are defined as gexels (gene expression elements), analogous to the pixels that constitute digital images. Similar to other computational tools dedicated to SrT data processing, MULTILAYER corrects for differences in total read counts per gexel, as such variations are considered to be artifactual (normalization; "Norm"). However, contrary to classical strategies that apply dimensionality reduction and clustering methodologies, normalized matrices are used by MULTILAYER for the computing of differential gene expression values relative to those for the average expression over the entire tissue. Similar to digital image processing, an agglomerative strategy is applied to reveal gene patterns defined by contiguous gexels, which are then compared to reveal spatial gene co-expression patterns expected to host functionally relevant information. A global comparison of all gene co-expression patterns leads to the partitioning of the initial spatial transcriptomics map into functionally relevant spatial community regions, which strongly improves the rendering (and its related biological coherence) relative to the spatial overlay of the identified clusters from dimensionality reduction strategies.

Inspired by digital image processing, which relies on contiguous pixel aggregation, we describe MULTILAYER, an analytical strategy dedicated to the processing of SrT readouts by pattern recognition from contiguous local transcriptomes. Such captured local transcriptomes are defined herein as gexels (gene expression elements), analogous to pixels, commonly described as units composing raster images in digital imaging. Hence, MULTILAYER processes SrT maps as a digital image, in which gexel patterns resulting from agglomerative clustering make it possible to highlight biologically relevant tissue substructures.

In this study, we compare the performance of MULTILAYER for the detection of differential gene expression with that of statistical methods previously applied for the analysis of SrT data. Then, we demonstrate its capacity to infer biologically relevant tissue substructures from the implemented gexel pattern recognition module and its enhanced performance relative to dimensionality reduction and clustering strategies on various SrT data, including the anatomical tissue stratification of a whole mouse embryo (DBiT-seq) (Liu et al., 2020). Finally, we present the performance of MULTILAYER for the analysis of high-resolution SrT data (Slide-seq; Rodriques et al., 2019) as well as for processing multiple datasets through an automated comparative batch module, anticipating the future needs of this rapidly evolving field. MULTILAYER is freely available as a stand-alone computational tool (https://github.com/SysFate/MULTILAYER)

## RESULTS

### Normalization, differential gene expression, detection of gene co-expression patterns, and digital tissue partitioning of SrT data performed by multilayer

MULTILAYER receives SrT matrices composed of spatial coordinates and read counts per gene as input. These matrices are converted into a grid view on which each spatial coordinate is associated with a gene expression element (or gexel), composed of read counts per gene for the local transcriptome. Raw SrT maps show variable total read counts per gexel, potentially due to technical issues during sample preparation (e.g., uneven tissue permeabilization, mRNA capture, etc). MULTILAYER applies quantile normalization (Hansen et al., 2012) across the gexels to address this problem, generating a uniform total read-count map over the entire grid. Other computational tools dedicated to SrT data processing, such as STviewer (Fernández Navarro et al., 2019), share these primary steps, including read-count normalization, prior application of dimensionality reduction, and clustering methodologies (Figure 1).

Under the assumption that the digital tissue map under study is nonhomogeneous, we aimed to infer changes in gene expression in a spatial context. MULTILAYER thus computes gene expression levels per gexel relative to the average gene expression within the tissue. Analogous to the terminology used in "bulk" gene expression analysis, we describe herein regions

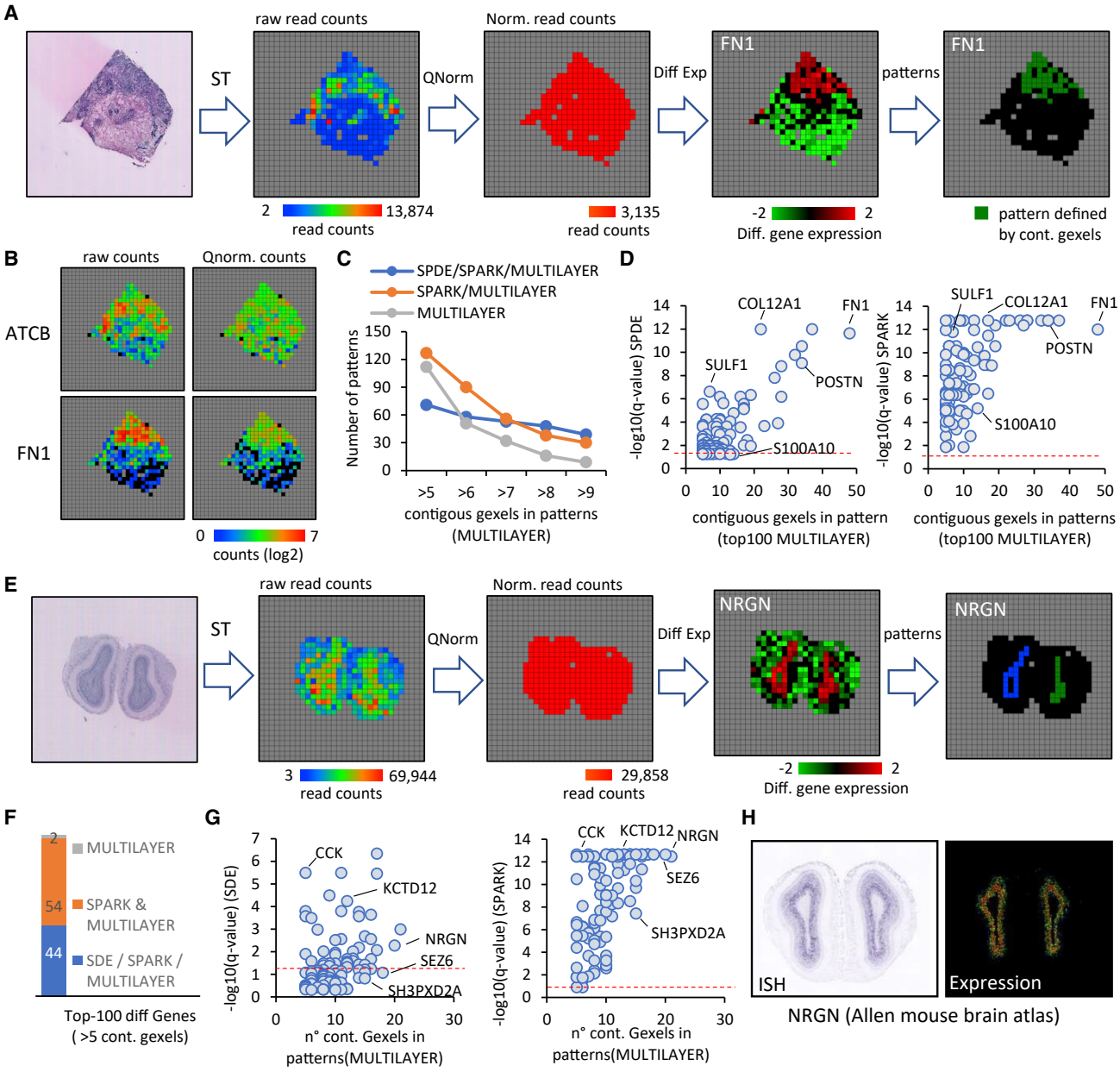



**Figure 2. The performance of MULTILAYER relative to two statistical approaches dedicated to the detection of spatial gene-expression patterns**

(A) Hematoxylin and eosin staining of breast cancer tissue used for a spatial transcriptomics assay (Ståhl et al., 2016) (left panel) and the processing of its corresponding data by MULTILAYER.

(B) Comparison between raw and quantile-normalized (Qnorm) read-count maps for the housekeeping gene ACTB and fibronectin 1 (FN1), known to be over-expressed in breast cancer (Wang et al., 2018).

(C) Number of common overexpressed gene patterns (false discovery rate [FDR] < 0.05) detected by SPARK (Sun et al., 2020), SpatialDE (SPDE; Svensson et al., 2018), and/or MULTILAYER relative to the number of contiguous gexels per pattern.

(D) Scatter plot displaying the statistical confidence associated with the top-100 genes (SPDE FDR ranking) inferred either by SPDE (left panel) or SPARK (right panel) relative to the number of contiguous gexels in patterns revealed by MULTILAYER. The FDR threshold (0.05) is highlighted by the red dashed line. Example of four genes (COL12A1, FN1, SULF1, and POSTN) reported to be highly significantly expressed by SPARK and SPDE, whereas S100A10 is above the defined FDR threshold by the SPARK statistical analysis only.

(E) Hematoxylin and eosin staining of mouse olfactory bulb (MOB) tissue used for a spatial transcriptomics assay (Ståhl et al., 2016) (left panel) and reprocessing of its corresponding data by MULTILAYER.

(F) Number of common significant gene patterns (FDR < 0.05) detected by SPARK (Sun et al., 2020), SPDE (Svensson et al., 2018), and/or MULTILAYER within the top-100 differentially expressed MOB genes (ranked by FDR predicted by SDE).

*(legend continued on next page)*

as upregulated or downregulated when the normalized read counts per gene are above or below the stated average behavior, respectively (Figure 1). Although this analysis is performed per gexel, MULTILAYER ranks differentially expressed genes based on the number of related gexels, hence providing a rapid view of genes that are overrepresented on the digital map on the basis of their relative expression (Figure S1B).

Similar to contextual classification strategies used for image analysis from pixel information (Toussaint, 1978), MULTILAYER detects gene expression patterns using an agglomerative strategy applied to contiguous gexels. This module generates a cleaner view of overexpressed genes within the tissue (Figure 1). Furthermore, it allows the identification of multiple patterns for the same overexpressed gene within the tissue, which per se is lost in all other strategies that rely on aggregating independent gexels by applying, for example, the t-distributed stochastic neighbor embedding (t-SNE) method (Figures S3, S4, and S7).

Having detected patterns for all overexpressed genes, MULTILAYER compares their spatial localization to infer their degree of co-expression (implementation of the Tanimoto and Dice similarity coefficient, see STAR methods). MULTILAYER expands this analysis through the ensemble of overexpressed genes to generate a graph in which nodes correspond to overexpressed genes within the tissue and edges their similarity coefficient, reflecting their degree of spatial co-expression (Figure S4). By applying the Louvain methodology for community detection (Blondel et al., 2008), MULTILAYER partitions the digital tissue map into biologically relevant tissue substructures. Using the aforementioned strategy, tissue partitioning performed by MULTILAYER surpasses the outcome obtained using the classical dimensionality reduction and clustering workflow, as illustrated in this study (Figures 1, 3, and S7).

## MULTILAYER detects differentially expressed genes with performance similar to existing methods

We validated the performance of the normalization and the detection of differential gene expression relative to that of existing methods using two public SrT datasets, one for the mouse olfactory bulb (MOB) and the other a human breast cancer tissue section (Ståhl et al., 2016). SrT landscaping over the breast cancer section gave rise to a digital map showing total counts per gexel varying between 2 and 13,874 reads, with an uneven distribution of high-count levels on one side of the tissue (Figure 2A). Such uneven distribution is also observed for the read counts associated to housekeeping gene actin beta (ACTB) (Figure 2B). Quantile normalization uniformized the total counts per gexel to 3,135 reads and rendered the ACTB read counts evenly distributed over the tissue. Quantile normalization did not affect the spatial read-count distribution associated with the gene fibronectin 1 (FN1), known to be overexpressed in breast cancer (Wang et al., 2018), which led to its detection as a spatially overexpressed gene with a pattern composed of several contiguous gexels (Figures 2A and 2B). Similarly, quantile normalization applied to the MOB dataset corrected for high (and unevenly spatially distributed) read-count levels associated with the housekeeping genes GAPDH and ACTB (Figure S1E). In the absence of such read-count correction, these genes would be considered to be spatially overexpressed, as illustrated by the artifactual pattern revealed by MULTILAYER. On the contrary, the spatial overexpression signature associated with the gene neurogranin (NRGN) was unchanged by the quantile treatment, as supported by the pattern detection performed by MULTILAYER, in agreement with publicly available *in situ* hybridization data (Figures 2E and S1A).

We compared the performance of MULTILAYER for the detection of differential gene expression using two existing statistical methods, SPARK (Sun et al., 2020) and SpatialDE (SPDE) (Svensson et al., 2018), previously used for the analysis of the breast cancer tissue and MOB data. For the breast cancer data, MULTILAYER detected >68% of all differentially expressed genes reported by SPARK (198 of the 290 genes, with a false discovery rate [FDR] < 0.05; Sun et al., 2020), from which 71 were in common with those reported by SPDE (from a total of 115) (Figure 2C). In contrast to the common genes detected between SPARK and MULTILAYER or by MULTILAYER alone, the fraction of commonly detected genes between SPDE, SPARK, and MULTILAYER showed a minor decrease when evaluated between 6 to 10 contiguous gexels (Figure 2C). This observation suggests that the differentially expressed genes commonly detected by these three methods are associated with patterns represented by a large number of contiguous gexels, a fact that was confirmed when comparing the top-100 SPDE-ranked genes with the number of contiguous gexels in patterns revealed by MULTILAYER (Figure 2D). SPDE produced conservative confidence descriptors, as illustrated by the associated q values for the overexpressed genes sulfatase-1 (SULF1), periostin (POSTN), and the member of the S100 protein family S100A10. In all cases, SPARK presented higher q values, in agreement with their rank assigned by MULTILAYER on the basis of the contiguous gexels within the detected pattern (Figure S2A). Finally, the fraction of differentially expressed genes detected by MULTILAYER alone corresponded to patterns composed of few contiguous gexels (112 genes with >5 contiguous gexels, and only 9 with >9 gexels, Figures 2C and S2B). Despite such behavior, these genes are enriched for cancer-related terms (DisGeNET association analysis; Figure S2D) and among them, factors such as cytochrome *c* oxidase subunit 6C (COX6C) or mucin-like 1 (MUCL1), which have been reported to be overexpressed in breast cancer, were retrieved (Figure S2C) (Chang et al., 2017; Conley et al., 2016).

In the case of the MOB data, MULTILAYER detected 98 of the top-100 SPDE-ranked genes (q value ranking, of which the top 67 had an FDR < 0.05; Sun et al., 2020), among which 44 were in common between all three methods (Figure 2F). The conservative statistical behavior of SPDE is highlighted when comparing the number of contiguous gexels per pattern detected by

(G) Same as (D) for the MOB dataset. Example of three genes (CCK, KCTD12, and NRGN) reported to be significantly expressed by SPARK and SPDE and two others (SEZ6 and SH3PXD2A) for which their associated confidence is beyond the defined FDR threshold by the SPARK statistical analysis only.
(H) *In situ* hybridization (ISH) and gene-expression data (Allen mouse brain atlas) for the neurogranin (NRGN) gene, revealing its spatial signature, which is coherent with the digitized view generated by MULTILAYER (displayed in E). The hematoxylin and eosin staining images reproduced in (A) and (E) were obtained from https://www.spatialresearch.org/.

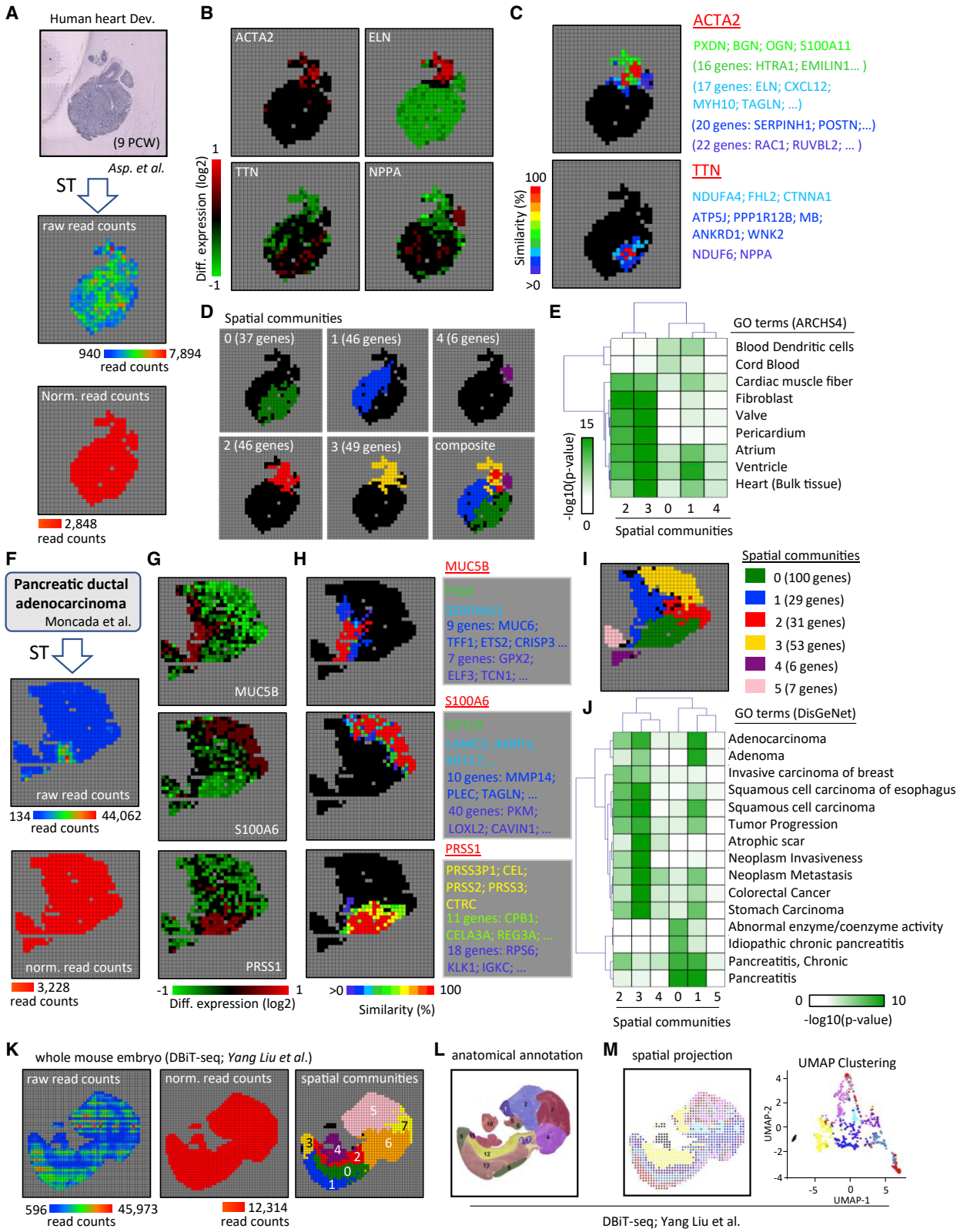



**Figure 3. MULTILAYER reveals functionally relevant tissue substructures by agglomerative clustering of digitized molecular signatures**

(A) MULTILAYER normalization of spatial transcriptomics data from developing human heart tissue (9 weeks post conception; Asp et al., 2019).

(B) Differential gene-expression signature (relative to the average behavior within the tissue) for ACTA2, ELN, TTN, and NPPA.

*(legend continued on next page)*

MULTILAYER with their corresponding q value (Figure 2G), notably by the low confidence assigned by SPDE to the genes SEZ6 and SH3PXD2A, but significantly detected by MULTILAYER and SPARK (Figure S2E).

Overall, this comparative analysis using SPARK and SPDE shows that MULTILAYER allows retrieval of most of the significant genes revealed by these methods.

## MULTILAYER efficiently partitions digitized tissue maps into biologically relevant substructures

In contrast to existing methods, MULTILAYER not only reveals differentially expressed genes but also uses their spatial information to partition digitized tissues into potentially relevant functional substructures. We processed public SrT maps for a variety of tissues, including human developing heart samples (Asp et al., 2019), pancreatic tumors (Moncada et al., 2020), as well as the recent processing of a whole mouse embryo (Liu et al., 2020), to assess its performance. As part of the heart study, MULTILAYER was instrumental in revealing the complexity of the tissue within 19 digital maps covering three described developmental stages (4.5–5, 6.5, and 9 post-conception weeks (PCW); Figure S5). We focused our attention on a tissue section collected at 9 PCW to highlight the performance of MULTILAYER (Figure S5, section 15). After applying quantile normalization to the raw read counts (Figure 3A), MULTILAYER detected several overexpressed genes, including those encoding the smooth muscle actin ACTA2, one of the components of the elastic fibers (elastin and ELN), the large abundant protein comprising the striated muscle Titin (TTN), and the natriuretic peptide NPPA (Figure 3B). The spatial gene overexpression signature for ACTA2 appeared to be concomitant with that of ELN and distinct from those observed for TTN and NPPA. This observation was confirmed by the spatial gene co-expression analysis performed by MULTILAYER, showing that ACTA2 and ELN present a similarity index >30% (Tanimoto distance) (Figure 3C), also observed for other factors, such as PXDN, BGN, S100A11, HTRA1, EMILIN1, CXCL12, MYH10, and TAGLN, some previously described to be expressed in the heart valve (Hinton et al., 2010; Munjal et al., 2014; Regalado et al., 2015). In a similar manner, TTN presented a co-expression signature with NPPA, as well as several other factors, such as NDUFA4, FHL2, and CTNNA1, known to present a specific left ventricle overexpression (as documented on the genotype-tissue expression portal (Lonsdale et al., 2013)(GTEx Consortium, 2020).

By extending the gene co-expression pattern detection over the entire tissue and applying spatial community partitioning, MULTILAYER revealed the presence of five communities that can be summarized within three major distinct tissue substructures (Figure 3D). Gexel communities "0" and "1" corresponded to two distinct regions, associated with the left and right ventricle and atrium of the heart (Figure 3E). In contrast, communities "2" and "3" showed redundant spatial location (Figure 3D) functionally related to pericardial tissue and the heart valve (Figure 3E). It is noteworthy that the aforementioned tissue substructures revealed by MULTILAYER, provides a clearer tissue partitioning than that obtained with dimensionality reduction and clustering strategies (Figure S7; Asp et al., 2019).

We analyzed pancreatic ductal adenocarcinoma SrT maps (Moncada et al., 2020) to further illustrate the performance of MULTILAYER on other types of tissue. After read-count normalization (Figure 3F), MULTILAYER detected a variety of spatially overexpressed genes, among them the mucin family member MUC5B, S100 calcium-binding protein A6 (S100A6), and the serine protease PRSS1 (Figure 3G). These three overexpressed genes showed a completely distinct spatial behavior, further confirmed by their gene co-expression patterns, as inferred by MULTILAYER (Figure 3H). Such distinct spatial patterns are in agreement with their previously described functional role, as MUC5B has been shown to be overexpressed in pancreatic ducts (Ringel and Löhr, 2003), S100A6 associated with pancreatic cancer development (Ohuchida et al., 2005), and PRSS1 expressed in normal pancreatic tissue, as it codes for trypsinogen, the enzyme secreted by this organ. Beyond these three distinct regions, MULTILAYER inferred up to six gexel communities (Figures 3I and S6), which can be summarized in four potentially functional relevant regions. Gexel community "0," was associated with functional terms such as pancreatitis or abnormal enzyme activity, most likely due to the fact that mutations of PRSS1 have been shown to be associated with hereditary pancreatic disorders (Shelton et al., 1993). This spatial region has been characterized as "normal pancreas tissue" as part of the histological annotation described by Moncada et al. (2020),

(C) Spatial gene co-expression analysis for ACTA2 and TTN. Gexels colored in red correspond to the location of the target genes (ACTA2 or TTN) and the other colored gexels show their co-expression pattern (Tanimoto similarity index).
(D) Spatial community tissue stratification from gene co-expression analysis and agglomerative clustering performed over the entire tissue and all overexpressed genes. The developing human heart tissue map was stratified into five spatial communities (from "0" to "4"), of which two show a highly redundant spatial localization pattern ("2" and "3").
(E) Gene ontology analysis (ARCSH4 tissue database; Lachmann et al., 2018) for each of the spatial communities retrieved within the developing human heart tissue, as performed by MULTILAYER. (F) MULTILAYER normalization of spatial transcriptomics data from pancreatic ductal adenocarcinoma tissue (Moncada et al., 2020).
(G) Differential gene-expression signature for MUC5B, S100A6, and PRSS1.
(H) Spatial gene co-expression analysis for MUC5B, S100A6, and PRSS1. Gexels colored in red correspond to the location of the target genes and the other colored gexels show their co-expression pattern (Tanimoto similarity index).
(I) Composite view of the six spatial communities detected in the pancreatic adenocarcinoma tissue from a gene co-expression analysis performed over the entire tissue and all overexpressed genes.
(J) Gene ontology analysis (DisGeNET database; Piñero et al., 2020) for each of the spatial communities displayed in (I), as performed by MULTILAYER.
(K) MULTILAYER processing of DBiT-seq data (deterministic bar coding in tissue for spatial omics sequencing; Liu et al., 2020) from a whole mouse embryo (50-μ gexel resolution). The right panel shows eight spatial communities revealed by agglomerative clustering of co-expressed gexels (patterns with >5 contiguous gexels; Tanimoto similarity index >10%).
(L) Anatomical annotation elaborated by Liu et al. (2020).
(M) UMAP clustering (right) and its corresponding spatial projection (left) performed by Liu et al. (2020) (images reproduced with authorization).

in agreement with the MULTILAYER functional annotation, which was devoid of tumor-related terms (Figure 3J). On the contrary, gexel community "3" was strongly associated with disease terms such as "adenocarcinoma," "tumor progression," and "neoplasm metastasis," in agreement with the histological annotation described by Moncada et al. (2020). Similarly, gexel communities "2" and "1" were associated with "cancer-related terms," but with a lower confidence, in agreement with the aforementioned histological differences (gexel community "1" described as duct epithelium and "2" as stroma; Moncada et al., 2020).

Finally, to further demonstrate the relevance of gexel community stratification, we processed data issued from the recently described DBiT-seq methodology, for which the authors used microfluidic channels to generate a spatial transcriptome map over a whole mouse embryo (Liu et al., 2020). The corresponding transcriptome map (50-μm resolution) is composed of gexels with total counts varying between ~600 and ~46,000 reads unevenly distributed throughout the tissue (Figure 3K). MULTILAYER uniformized the total counts per gexel to ~12,000 reads and allowed to reveal up to eight distinct spatial communities (Figures 3K and S7I). The spatial localization of the inferred communities showed a strong correlation with the anatomical annotation presented within the DBiT-seq study (Figures 3L and S7J; Liu et al., 2020), better than the unsupervised UMAP clustering (Figure 3M) or a similar dimensionality reduction that can be obtained using the t-SNE algorithm (Figure S7J). This enhanced MULTILAYER performance relative to other strategies that rely on aggregating independent gexels by applying the t-SNE algorithm was also verified for the developing human heart and the pancreatic ductal adenocarcinoma datasets (Figures S7A–S7F).

Overall, MULTILAYER made it possible to perform automated tissue stratification, showing biologically relevant roles coherent with the findings revealed in the studies from which we collected the data. It is noteworthy that, contrary to previous studies, MULTILAYER partitions the digital map based on contiguous gexel information and provides ranked lists of overexpressed genes and relevant gene co-expression patterns to the user, thus enhancing the molecular characterization of the spatial information in a self-supervised manner, similar to that used for tissue image segmentation (reviewed in Gildenblat and Klaiman [2019]).

## MULTILAYER allows the processing of high-resolution SrT maps by incorporating a super-gexel agglomerative compression module

Most of the current available SrT maps come from glass slides on which barcoded-polyT DNA probes are printed. The manufacturing constraints of such DNA arrays provide a resolution of ~100 μm (equivalent to ~10–40 cells per gexel), with a number of spots ranging between ~1,000 and ~5,000 (when considering the recent commercial upgrade of the original SrT protocol), covering a surface of ~6 × 6 mm (Ståhl et al., 2016). An alternative strategy, based on the use of uniquely DNA-barcoded beads deposited onto a glass coverslip, enhanced the SrT resolution to 10 μm. This methodology, known as Slide-seq, allowed the generation of high-resolution SrT maps within a circular surface of 3 mm in diameter containing ~70,000 uniquely DNA-barcoded beads (Rodriques et al., 2019).

Aiming to use MULTILAYER to analyze high-resolution Slide-seq maps but concerned by the technical constraints related to (1) the low number of read counts per gene retrieved within the gexels (Figure S8) and (2) the high number of gexels within the SrT map impacting the computational performance (including the display functionalities); we have implemented a complementary script allowing to reduce the SrT map complexity. This ad hoc module, called "MULTILAYER compressor," generates super-gexels by agglomerating contiguous gexels defined by a user-provided compression factor. This approach, previously described for image-segmentation strategies (Stutz et al., 2018), allowed enhancement of the number of counts per gene within super-gexels to levels even comparable to those retrieved in regular SrT maps (Figure S8). Furthermore, it reduced the computational requirements, such that MULTILAYER was able to deconvolve the potentially functional relevant complexity of the digital tissue. This last aspect has been highlighted by the analysis of public Slide-seq data, including mouse hippocampus and sagittal cortex maps (Fan et al., 2020; Rodriques et al., 2019). In both cases, raw Slide-seq maps composed of ~ 70,000 high-resolution gexels were compressed by factors of 60×, 100×, and 175×, leading to grid sizes compatible with the performance of MULTILAYER (Figures 4 and S8–S11). Normalization and differential gene expression analysis performed on the hippocampus map (reduction factor of 60×) allowed us to identify spatially distinct overexpression signatures for factors such as protein phosphatase 3 catalytic subunit alpha (PPP3CA), Purkinje cell protein 4 (PCP4), and synaptosome-associated protein 25 (SNAP25) (Figures 4B and S9). This was further supported by global tissue stratification from a comparison of gene co-expression patterns retrieved over the entire tissue (43 spatial communities; Figure 4C). The use of higher compression factors (100× and 175×) did not affect the observed overexpression signature for PPP3CA, PCP4, or SNAP25 and increased their related read counts and differential overexpression levels, in agreement with the agglomerative strategy used for generating super-gexels (Figures 4B and S9C). Furthermore, it reduced the number of detected spatial gexel communities (13 and 8 spatial communities for 100× and 175×, respectively) but retained the global digital tissue substructures (Figures 4C, 4E, and 4G). Finally, the relevance of the various spatial gexel communities were confirmed by gene ontology enrichment analysis (ARCHS4 tissue database; Lachmann et al., 2018) assessed using the three described compression factors (Figures 4D, 4F, and 4H).

A similar analysis performed on the cortex map (reduction factor of 60×) allowed us to identify spatially distinct overexpression signatures for factors such as the gene transthyretin (TTR) and calcium/calmodulin-dependent protein kinase II inhibitor 1 (CAMK2N1) (Figure S10). We confirmed their gene expression relevance within the mouse cortex by gene ontology term analysis performed over the spatially co-expressed genes at various digital compression factors (60×, 100×, and 175×) (Figures S10E–S10I). The extension of the co-expression pattern over all genes within the tissue map allowed us to infer >40 super-gexel communities on the cortex map using the 60× compression factor (Figure S11A). Gene ontology analysis performed by MULTILAYER (ARCHS4 tissue database) showed the enrichment of terms such as cerebral cortex, superior frontal gyrus,

CellPress
OPEN ACCESS

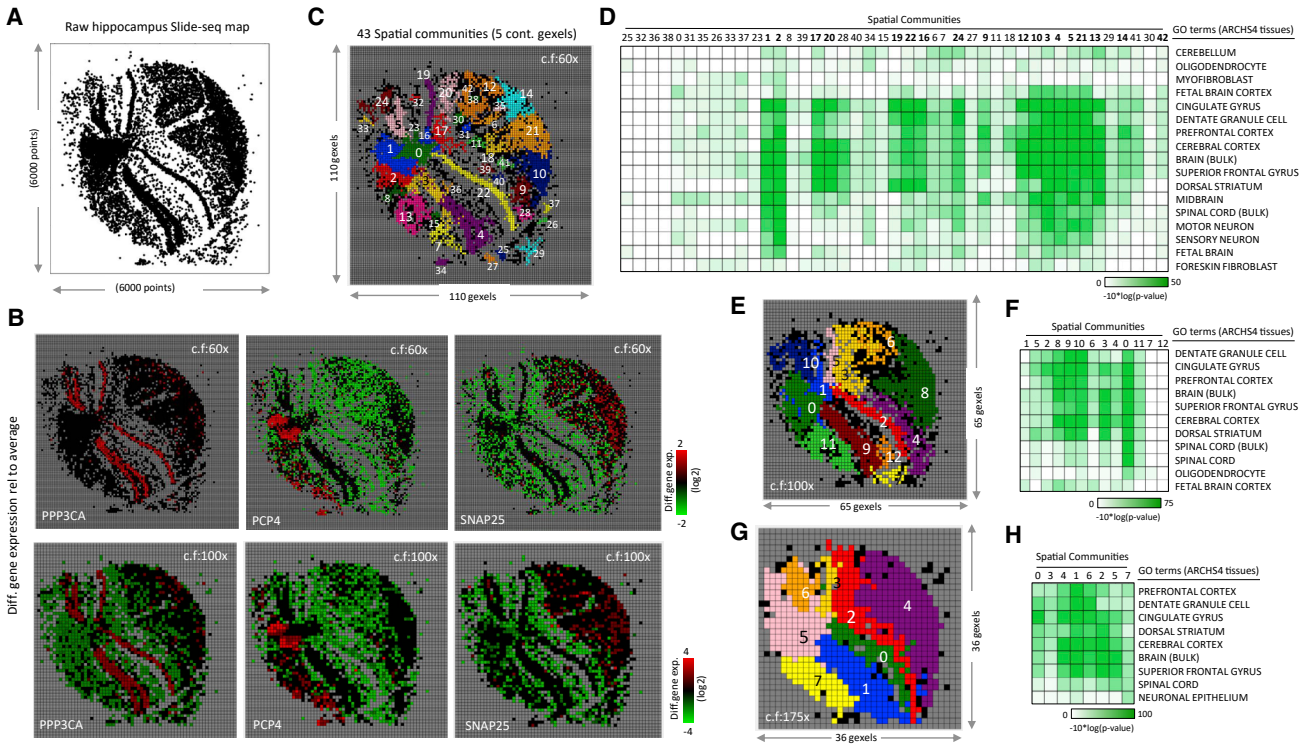



**Figure 4. High-resolution hippocampus spatial transcriptomics map analyzed by MULTILAYER**

(A) Raw hippocampus Slide-seq map displaying the presence of at least one read count per position. (B) Differential spatial expression signatures associated with the factors PPP3CA, PCP4, and SNAP25 after applying compression factors (c.f.) of 60× and 100×, respectively.

(C) Spatial communities revealed on the hippocampus ST map after applying a compression factor of 60×. MULTILAYER compressor reduced the complexity of the original map (displayed in A) to a grid composed of 110 × 110 super-gexels, which was then processed by MULTILAYER.

(D) Gene ontology enrichment analysis performed on the 43 spatial communities displayed in (C).

(E and G) Spatial communities revealed on the hippocampus ST map after applying compression factors (c.f.) of 100× and 175×, respectively.

(F and H) Gene ontology enrichment analysis performed on the spatial communities displayed in (E) and (G), respectively. ARCHS4 tissues: GO terms database from massive mining of publicly available RNA-seq data from human and mouse (Lachmann et al., 2018).

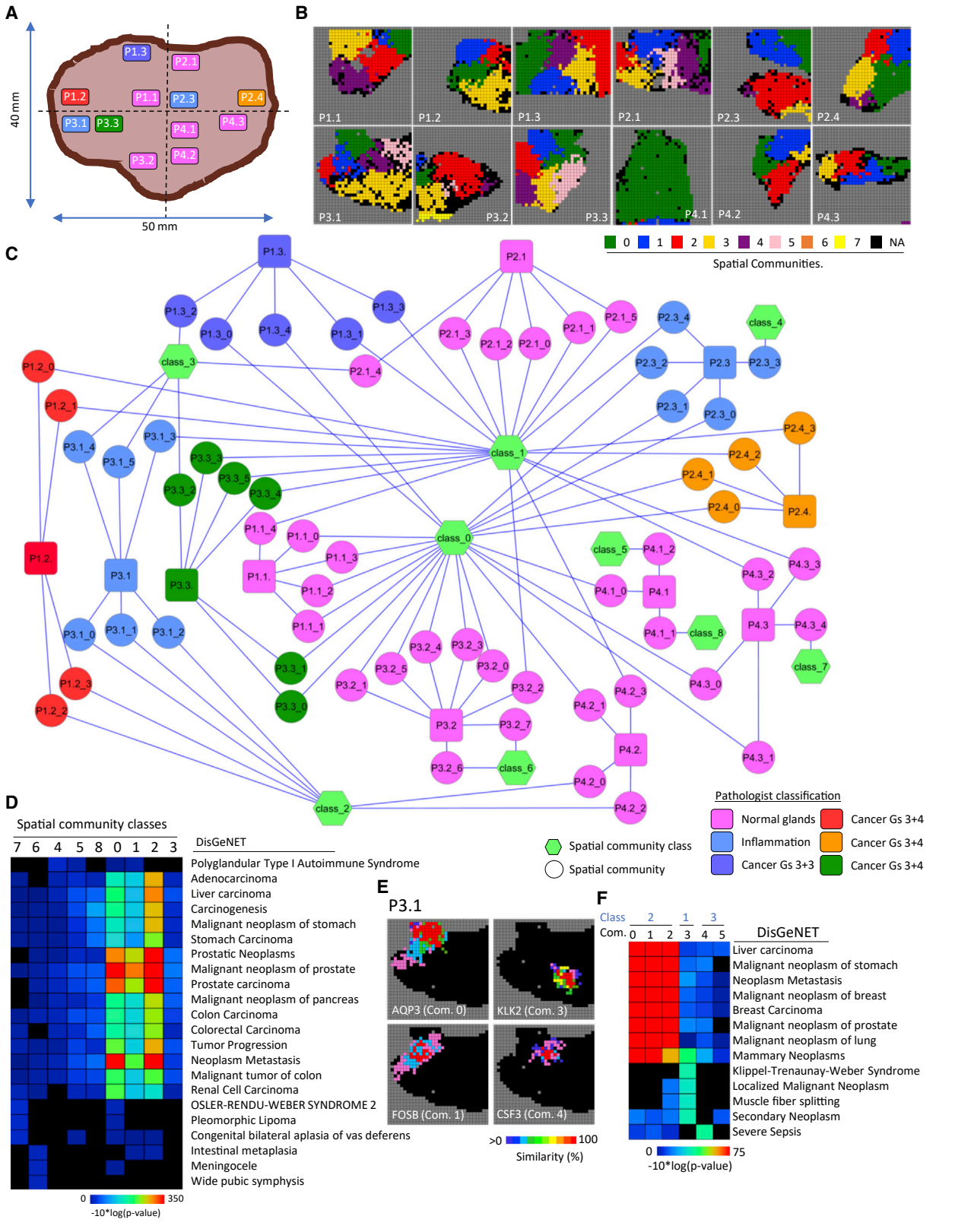
dentate granule cell, motor neuron, neuronal epithelium, dorsal striatum, and spinal cord (Figure S11D). The use of a compression factor of 100× reduced the digital tissue stratification to 15 communities (Figure S11B) and down to seven when a 175× compression factor was applied (Figure S11C). In both cases, the major spatial tissue stratification remained visible, further supported by their associated gene ontology terms (Figures S11E and S11F).

Overall, MULTILAYER allowed the stratification of high-resolution but sparse SrT maps by the use of a super-gexel agglomerative compression strategy.

## MULTILAYER provides an enhanced tissue heterogeneity classification when comparing multiple digitized tissue maps

A major question to address when analyzing multiple tissues from related samples is whether their inferred substructures (herein referred as spatial communities) share commonalities and differences that could enhance our understanding of their molecular inter-relationship. Recently, Berglund et al. generated SrT maps from 12 spatially separated biopsies from a cancerous prostate, for which a pathological annotation, based on a histological analysis (Gleason Grading), was performed (Figure 5A) (Berglund et al., 2018). We addressed this question by implementing a "batch mode" within MULTILAYER, allowing the processing of multiple SrT maps, and their comparison on the basis of their stratified spatial communities. MULTILAYER inferred spatial community substructures within all 12 sections and revealed their significantly enriched disease-gene associations (Figures 5B and S12).

MULTILAYER performed inter-tissue comparisons by constructing a graph in which spatial communities per tissue were associated with their relevant gene co-expression patterns. Such inter-tissue graph was partitioned (Louvain methodology Blondel et al., 2008) into nine "classes," corresponding to the relationship between tissue substructures retrieved among all 12 biopsies (Figure 5C). Noteworthy, MULTILAYER's partitioning revealed that all tissues present molecular signatures related to prostate cancer progression on at least one substructure, regardless of the histological classification (Figures 5C, 5D, and S12). For example, tissue biopsies histologically classified as "normal glands" (P1.1, P2.1, P3.2, P4.1, P4.2, and P4.3) showed gene co-expression patterns associated with factors such as the membrane cell-junction protein claudin-4 (CLDN4), known to be overexpressed in primary and metastatic prostate cancer (Landers et al., 2008); growth/differentiation




*(legend on next page)*

factor-15 (GDF-15), the overexpression of which has been associated with prostate cancer progression (Vaňhara et al., 2012); the gene ACPP, encoding prostatic acid phosphatase, associated with prostatic hyperplasia and also observed in prostate carcinoma (as shown in the human protein atlas database); kallikrein-related peptidase 2 (KLK2), encoding a trypsin-like serine protease primarily expressed in the prostate, the overexpression of which is considered to be a prognostic marker for prostate cancer risk (Shang et al., 2014); and activating transcription factor 3 (ATF3), shown to be upregulated in oncogenic stress and described as a tumor suppressor response, presenting an inhibitory effect on androgen-receptor signaling (Wang et al., 2015) (Figures S12–S14). Similarly, tissue P3.1, classified as ''inflammation,'' was stratified in six spatial community substructures, four of which were associated with inter-tissue classes functionally enriched for cancer-related terms (class 2: communities 0, 1, and 2; class 1: community 3) (DisGeNET (Piñero et al., 2020) analysis; Figures 5C, 5D, and 5F). This annotation was supported by gene co-expression patterns related to factors such as the Fos protein FOSB, known to form transcriptionally active heterodimers with Jun proteins and reported to be overexpressed in prostate cancer cell lines (Barrett et al., 2017), as well as in prostate cancer biopsies (Berglund et al., 2018), or KLK2 (Figure 5E). Furthermore, although the ''inflammation'' classification was supported by the local overexpression of the gene aquaporin-3 (AQP3; community ''0'') (Figure 5E), the gene co-expression analysis for this factor revealed the presence of other players within the same community, including serine peptidase inhibitor kazal type 1 (SPINK1), previously described as a marker for a molecular subtype of prostate cancer (Johnson et al., 2016) (Figures S12–S14). Finally, the spatial communities ''4'' and ''5'' appeared to be devoid of major cancer-related terms (supporting their association with class ''3'') but still showed molecular signatures related to prostate cancer incidence, such as sever sepsis (Figure 5F). Indeed, community ''4'' showed a gene co-expression signature related to colony-stimulating factor 3 (CSF3), known to regulate the generation of infection-protective granulocytes and macrophages (Metcalf, 2010), an aspect that supports the histological classification of this tissue as ''inflammation.''

Finally, the tissue biopsies histologically classified as cancer did not systematically present all spatial communities related to cancer-related terms. Certain community substructures in tissues P1.3 (histologically classified as ''inflammation'') and P3.3 1 (histologically classified as ''normal glands'') were associated with class ''3,'' further supporting the necessity of molecular tissue stratification to better define tumor progression.

## DISCUSSION

Although the use of single-cell transcriptomics for studying the molecular complexity of tissues is gaining popularity, spatial transcriptomics strategies are anticipated to take over in the following years, thanks to the efforts to democratize access to the required physical supports. Indeed, although SrT is systematically considered to be a ''non-single-cell resolution'' assay, in reality, all single-cell ''omics'' approaches converge to aggregate multiple cell readouts into clusters to infer their potentially functional relevance. Similarly, most of the computational algorithms applied to SrT maps process gexels as independent units (i.e., by applying clustering strategies classically used in single-cell ''omics'' assays), thus underexploring the available spatial information.

Here, we considered SrT data as a digital image composed of an ensemble of gexels arranged in a two-dimensional grid. Hence, we demonstrated that the principles applied to digital image processing, which relies on contiguous pixel aggregation, can help to reveal biologically relevant tissue substructures. Specifically, the proposed stand-alone package MULTILAYER allows the ''rationalization'' of spatial information by analyzing the presence of contiguous gexels showing the same gene expression behavior (herein defined as a gene pattern), leading to the stratification of the digital map into molecular tissue substructures. Indeed, in contrary to other methodologies, in which gexels are considered to be independent units that converge into given clusters after dimensionality reduction processing, MULTILAYER captures gene expression patterns, which are then converged toward regions due to their spatial co-localization. We demonstrated within this study that this major conceptual strategy can lead to better resolved molecular tissue stratification, which can be explained by the fact that tissue substructures are expected to arise from the organization of contiguous cells sharing common cell fates, issued from defined gene programs.

MULTILAYER provides a self-supervised strategy for the processing of SrT maps. It highlights relevant overexpressed gene patterns, leading to spatial tissue partitioning, and infers their functionally relevant gene ontology associations. Due to its ''MULTILAYER'' architecture, it provides the means to process all types of SrT maps, including high-resolution data. Furthermore, it provides the means to compare multiple SrT maps, an aspect that is currently limited by the scarcity of studies

**Figure 5. MULTILAYER reveals an enhanced discrepancy between normal and cancer-related tissue sub-structures within multiple prostate cancer tissue biopsies**

(A) Scheme representing the spatial location of 12 tissue biopsies collected from a cancerous prostate and colored in agreement with the histological classification, as described by Berglund et al. (Berglund et al., 2018).

(B) Spatial transcriptome maps from the biopsies illustrated in (A) and processed by MULTILAYER to infer spatial community molecular substructures (color-coded, NA, non-assigned).

(C) Inter-tissue comparison performed by MULTILAYER (batch-mode) organizing all spatial communities (round nodes) into nine ''classes'' (green hexagonal nodes). In addition, the tissue biopsy of origin is displayed (rounded-square nodes). Nodes are colored in agreement with the histological classification described by Berglund et al. (218).

(D) Relevant gene-disease association inferred for the spatial community classes displayed in (C).

(E) Example of gene co-expression patterns detected in tissue P3.1 for different spatial communities. Gexels in red correspond to the query gene and the others to gene co-expression similarity patterns (Tanimoto index in percent).

(F) Relevant gene-disease association analysis for each of the spatial communities retrieved for tissue P3.1. The corresponding spatial community classes are also displayed (light blue). DisGeNET: disease-gene association discovery platform (Piñero et al., 2020), Gs, Gleason score for cancer staging; Com, community.

presenting the required data. However, it is anticipated that it will be of extreme interest, for example, in the context of developmental or molecular diagnostic studies.

Overall, we anticipate that MULTILAYER corresponds to the first version of algorithms dedicated to the processing of ''molecular tissues,'' which combined with other strategies, such as single-cell ''omics,'' may allow the reconstitution of digital maps of all organs within the human body, as well as contribute to the development of molecular diagnostic strategies in the future in personalized medicine.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Normalization
  - Spatial differential gene expression
  - Gene-expression pattern detection
  - Gene co-expression patterns similarity
  - Identification of tissue communities by gene co-expression-pattern partitioning
  - Gene ontology analysis
  - MULTILAYER Compressor ad-hoc module
  - Comparison with SPARK and SpatialDE statistical methods
  - Data availability
  - Code availability

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2021.04.008.

## ACKNOWLEDGMENTS

This work was supported by Genopole Thematic Incentive Actions funding (ATIGE-2017) and the institutional bodies, the CEA, CNRS, and Université d'Évry – Val d'Essonne.

## AUTHOR CONTRIBUTIONS

Conceptualization, J.M. and M.A.M.P.; methodology, J.M., B.M., and M.A.M.P.; software development, J.M. and B.M.; scientific evaluation, B.M.C. and M.A.M.P. writing, review, and editing, J.M., B.M.C., and M.A.M.P.; funding acquisition, M.A.M.P.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 25, 2020
Revised: January 8, 2021
Accepted: April 13, 2021
Published: May 7, 2021

## REFERENCES

Asp, M., Giacomello, S., Larsson, L., Wu, C., Fürth, D., Qian, X., Wärdell, E., Custodio, J., Reimegård, J., Salmén, F., et al. (2019). A spatiotemporal organ-wide gene expression and cell atlas of the developing human heart. Cell *179*, 1647–1660.e19.

Barrett, C.S.X., Millena, A.C., and Khan, S.A. (2017). TGF-β effects on prostate cancer cell migration and invasion require FosB. Prostate *77*, 72–81.

Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. *37*, 38–44.

Bergenstråhle, J., Larsson, L., and Lundeberg, J. (2020). Seamless integration of image and molecular analysis for spatial transcriptomics workflows. BMC Genomics *21*, 482.

Berglund, E., Maaskola, J., Schultz, N., Friedrich, S., Marklund, M., Bergenstråhle, J., Tarish, F., Tanoglidi, A., Vickovic, S., Larsson, L., et al. (2018). Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. Nat. Commun. *9*, 2419.

Birnbaum, K.D. (2018). Power in numbers: single-cell RNA-seq strategies to dissect complex tissues. Annu. Rev. Genet. *52*, 203–221.

Blondel, V.D., Guillaume, J., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. *2008*, 10008.

Chang, F.W., Fan, H.C., Liu, J.M., Fan, T.P., Jing, J., Yang, C.L., and Hsu, R.J. (2017). Estrogen enhances the expression of the multidrug transporter gene ABCG2-increasing drug resistance of breast cancer cells through estrogen receptors. Int. J. Mol. Sci. *18*.

Conley, S.J., Bosco, E.E., Tice, D.A., Hollingsworth, R.E., Herbst, R., and Xiao, Z. (2016). HER2 drives Mucin-like 1 to control proliferation in breast cancer cells. Oncogene *35*, 4225–4234.

Fan, Z., Chen, R., and Chen, X. (2020). SpatialDB: a database for spatially resolved transcriptomes. Nucleic Acids Res. *48*, D233–D237.

Feng, C., Liu, S., Zhang, H., Guan, R., Li, D., Zhou, F., Liang, Y., and Feng, X. (2020). Dimension reduction and clustering models for single-cell RNA sequencing data: a comparative study. Int. J. Mol. Sci. *21*, 2419.

Fernández Navarro, J., Lundeberg, J., and Ståhl, P.L. (2019). ST viewer: a tool for analysis and visualization of spatial transcriptomics datasets. Bioinformatics *35*, 1058–1060.

Gildenblat, J., and Klaiman, E. (2019). Self-supervised similarity learning for digital pathology. arXiv https://arxiv.org/abs/1905.08139.

GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *11*, 1318–1330.

Hansen, K.D., Irizarry, R.A., and Wu, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics *13*, 204–216.

Hinton, R.B., Adelman-Brown, J., Witt, S., Krishnamurthy, V.K., Osinska, H., Sakthivel, B., James, J.F., Li, D.Y., Narmoneva, D.A., Mecham, R.P., and Benson, D.W. (2010). Elastin haploinsufficiency results in progressive aortic valve malformation and latent valve disease in a mouse model. Circ. Res. *107*, 549–557.

Johnson, M.H., Ross, A.E., Alshalalfa, M., Erho, N., Yousefi, K., Glavaris, S., Fedor, H., Han, M., Faraj, S.F., Bezerra, S.M., et al. (2016). SPINK1 defines a molecular subtype of prostate cancer in men with more rapid progression in an at risk, natural history radical prostatectomy cohort. J. Urol. *196*, 1436–1444.

Kiselev, V.Y., Andrews, T.S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. Nat. Rev. Genet. *20*, 273–282.

Lachmann, A., Torre, D., Keenan, A.B., Jagodnik, K.M., Lee, H.J., Wang, L., Silverstein, M.C., and Ma'ayan, A. (2018). Massive mining of publicly available RNA-seq data from human and mouse. Nat. Commun. *9*, 1366.

Landers, K.A., Samaratunga, H., Teng, L., Buck, M., Burger, M.J., Scells, B., Lavin, M.F., and Gardiner, R.A. (2008). Identification of claudin-4 as a marker highly overexpressed in both primary and metastatic prostate cancer. Br. J. Cancer *99*, 491–501.

Liu, Y., Yang, M., Deng, Y., Su, G., Enninful, A., Guo, C.C., Tebaldi, T., Zhang, D., Kim, D., Bai, Z., et al. (2020). High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. Cell *183*, 1665–1681.e18.

Lonsdale, J., Thomas, J., Salvatore, M., et al. (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet *45*, 5801318–5851330, https://doi.org/10.1038/ng.2653.

Metcalf, D. (2010). The colony-stimulating factors and cancer. Nat. Rev. Cancer *10*, 425–434.

Moncada, R., Barkley, D., Wagner, F., Chiodin, M., Devlin, J.C., Baron, M., Hajdu, C.H., Simeone, D.M., and Yanai, I. (2020). Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. Nat. Biotechnol. *38*, 333–342.

Munjal, C., Opoka, A.M., Osinska, H., James, J.F., Bressan, G.M., and Hinton, R.B. (2014). TGF-β mediates early angiogenesis and latent fibrosis in an Emilin1-deficient mouse model of aortic valve disease. Dis. Model. Mech. *7*, 987–996.

Ohuchida, K., Mizumoto, K., Ishikawa, N., Fujii, K., Konomi, H., Nagai, E., Yamaguchi, K., Tsuneyoshi, M., and Tanaka, M. (2005). The role of S100A6 in pancreatic cancer development and its clinical implication as a diagnostic marker and therapeutic target. Clin. Cancer Res. *11*, 7785–7793.

Piñero, J., Ramírez-Anguita, J.M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., and Furlong, L.I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. *48*, D845–D855.

Raimundo, F., Vallot, C., and Vert, J.P. (2020). Tuning parameters of dimensionality reduction methods for single-cell RNA-seq analysis. Genome Biol. *21*, 212.

Regalado, E.S., Guo, D.C., Prakash, S., Bensend, T.A., Flynn, K., Estrera, A., Safi, H., Liang, D., Hyland, J., Child, A., et al. (2015). Aortic disease presentation and outcome associated with ACTA2 mutations. Circ. Cardiovasc. Genet. *8*, 457–464.

Ringel, J., and Löhr, M. (2003). The MUC gene family: their role in diagnosis and early detection of pancreatic cancer. Mol. Cancer *2*, 9.

Rodriques, S.G., Stickels, R.R., Goeva, A., Martin, C.A., Murray, E., Vanderburg, C.R., Welch, J., Chen, L.M., Chen, F., and Macosko, E.Z. (2019). Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. Science *363*, 1463–1467.

Shang, Z., Niu, Y., Cai, Q., Chen, J., Tian, J., Yeh, S., Lai, K.P., and Chang, C. (2014). Human kallikrein 2 (KLK2) promotes prostate cancer cell growth via function as a modulator to promote the ARA70-enhanced androgen receptor transactivation. Tumour Biol. *35*, 1881–1890.

Shelton, C., Solomon, S., LaRusch, J., and Whitcomb, D.C. (1993). PRSS1-related hereditary pancreatitis. In GeneReviews, M.P. Adam, H.H. Ardinger, R.A. Pagon, S.E. Wallace, L.J. Bean, K. Stephens, and A. Amemiya, eds. (University of Washington) https://www.ncbi.nlm.nih.gov/books/NBK84399/.

Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science *353*, 78–82.

Stutz, D., Hermans, A., and Leibe, B. (2018). Superpixels: an evaluation of the state-of-the-art. Comput. Vision Image Underst. *166*, 1–27.

Sun, S., Zhu, J., Ma, Y., and Zhou, X. (2019). Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. Genome Biol. *20*, 269.

Sun, S., Zhu, J., and Zhou, X. (2020). Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nat. Methods *17*, 193–200.

Svensson, V., Teichmann, S.A., and Stegle, O. (2018). SpatialDE: identification of spatially variable genes. Nat. Methods *15*, 343–346.

Toussaint, G.T. (1978). The use of context in pattern recognition. Pattern Recognit. *10*, 189–204.

van den Brink, S.C., Sage, F., Vértesy, Á., Spanjaard, B., Peterson-Maduro, J., Baron, C.S., Robin, C., and van Oudenaarden, A. (2017). Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nat. Methods *14*, 935–936.

Vaňhara, P., Hampl, A., Kozubík, A., and Souček, K. (2012). Growth/differentiation factor-15: prostate cancer suppressor or promoter? Prostate Cancer Prostatic Dis. *15*, 320–328.

Wang, Y., Xu, H., Zhu, B., Qiu, Z., and Lin, Z. (2018). Systematic identification of the key candidate genes in breast cancer stroma. Cell. Mol. Biol. Lett. *23*, 44.

Wang, Z., Xu, D., Ding, H.F., Kim, J., Zhang, J., Hai, T., and Yan, C. (2015). Loss of ATF3 promotes Akt activation and prostate cancer development in a Pten knockout mouse model. Oncogene *34*, 4975–4984.

## STAR★METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Deposited data | | |
| Breast cancer SrT raw matrix (ST) | Ståhl et al., 2016 | https://github.com/SysFate/MULTILAYER/tree/master/Data/Breast_cancer |
| Development_heart raw matrices (ST) | Asp et al., 2019 | https://github.com/SysFate/MULTILAYER/tree/master/Data/Development_heart |
| Mouse_Olfactory_Bulb raw matrix (ST) | Ståhl et al., 2016 | https://github.com/SysFate/MULTILAYER/tree/master/Data/Mouse_Olfactory_Bulb |
| Pancreatic_adenocarcinoma raw matrix (ST) | Moncada et al., 2020 | https://github.com/SysFate/MULTILAYER/tree/master/Data/Pancreatic_adenocarcinoma |
| Prostate_cancer raw matrices (ST) | Berglund et al., 2018 | https://github.com/SysFate/MULTILAYER/tree/master/Data/Prostate_cancer |
| High_resolution_brain raw matrices (Slide-seq) | Rodriques et al., 2019 | https://github.com/SysFate/MULTILAYER/tree/master/Data/High_resolution_brain |
| Whole_mouse_embryo raw matrix (DBiT-Seq) | Liu et al., 2020 | https://github.com/SysFate/MULTILAYER/tree/master/Data/Whole_mouse_embryo |
| Spatial gene expression readouts for the mouse olfactory bulb (Rep11_MOB) and human breast-cancer tissue data (Layer2_BC) processed with SPARK | Kindly provided by Dr. Jiaqiang Zhu (Sun et al., 2020) | https://github.com/SysFate/MULTILAYER/tree/master/Data/SPARK_diffGenes |
| Spatial gene expression readouts for the mouse olfactory bulb (Rep11_MOB_spe.csv) and human breast-cancer tissue data (Layer2_BC_spe.csv) processed with Spatial DE | Sun et al., 2020 | https://github.com/xzhoulab/SPARK-Analysis/tree/master/output |
| Software and algorithms | | |
| MULTILAYER | This paper | https://github.com/SysFate/MULTILAYER |
| MULTILAYER compressor | This paper | https://github.com/SysFate/MULTILAYER |

### RESOURCE AVAILABILITY

#### Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Marco Antonio Mendoza-Parra (mmendoza@genoscope.cns.fr).

#### Materials availability

This study did not generate new materials.

#### Data and code availability

This paper analyzes existing, publicly available data. These datasets were converted into a uniform format compatible with MULTILAYER requirements and are provided in the key resources table.

MULTILAYER and MULTILAYER Compressor are available at https://github.com/SysFate/MULTILAYER.

Scripts used to generate the figures presented in this paper – other than MULTILAYER outputs - are not provided in this paper but are available from the Lead Contact on request.

Any additional information required to reproduce this work is available from the Lead Contact.

## METHOD DETAILS

### Normalization

MULTILAYER corrects for differences in total read counts within gexels, as such variations are considered to be technical artifacts arising from the sample preparation. Quantile normalization (previously described for correcting technical variations within RNA-seq assays (Hansen et al., 2012)) is applied as follows. A pseudo-count of 1 is added to all gexels to avoid handling null values. Read-counts per gene within gexels are sorted on the basis of their frequency and then the average read-counts across all ranked gexels is computed based on their rank (i.e., average within the highest, middle, or lowest values). Finally, the average read-counts are incorporated instead of the original counts and the read-count distribution reorganized as initially. As a consequence, a constant value is retrieved across all gexels when all read-counts per gexel are added after normalization, corresponding to an ideal situation in which all coordinates within the digitized tissue are composed of the same sequencing coverage level.

### Spatial differential gene expression

Based on the assumption that the tissue under study is not homogenous, MULTILAYER performs differential expression analysis to identify over/under-expressed genes relative to the global behavior within the tissue. The average of the read counts per gene within the tissue is computed and the read counts per gene within gexels then expressed relative to the average value ($log_2$). Differentially expressed genes within gexels are defined by a threshold value of two-fold (1 or -1 in $log_2$) as a default parameter. As part of the ''differential expression'' panel, MULTILAYER displays a ranking of induced or repressed genes based on the number of gexels within the tissue, allowing the intuitive identification of the most relevant overexpressed genes.

### Gene-expression pattern detection

As in digital image processing, MULTILAYER applies an iterative agglomerative strategy (sklearn.cluster.AgglomerativeClustering) over contiguous gexels associated with a given upregulated gene. At the end of the process, gene patterns showing a user-defined minimal number of contiguous gexels are retained for downstream processing (default threshold: 10 contiguous gexels). The parameters used in the agglomerative clustering method are the number of clusters (n_clusters): none, affinity: Euclidean, linkage: single, distance threshold (distance_threshold): 1.5.

### Gene co-expression patterns similarity

Once the gene patterns over the whole tissue have been detected, MULTILAYER compares their localization to assess their relevant spatial co-expression. Two similarity metrics are implemented in MULTILAYER: the Tanimoto/Jaccard and Dice/Sorensen similarity indices. Specifically, the gene co-expression pattern similarity is evaluated as follows:

$$\text{Tanimoto index} = G_A \cap G_B \,/\, G_A \cup G_B$$

$$\text{Dice index} = 2^{*} (G_A \cap G_B) \,/\, (G_A \cup G_B)$$

where $G_A$ and $G_B$ correspond to the number of gexels associated with Gene A and Gene B, respectively. All figures presented in this article were obtained using the Tanimoto/Jaccard similarity index.

Within the gene co-expression pattern panel of MULTILAYER, overexpressed gene patterns are ranked on the basis of the number of contiguous gexels. Furthermore, the gene co-expression similarity analysis displays gexels colored on the basis of their co-expression similarity index, allowing visualization of the extent of co-expressed patterns with the queried gene. In addition, MULTILAYER provides the possibility to perform a gene ontology enrichment analysis based on their inferred co-expressed genes (see below).

### Identification of tissue communities by gene co-expression-pattern partitioning

Gene co-expression patterns detected over the whole tissue are represented within MULTILAYER as a graph composed of nodes representing the assessed gene patterns and edges revealing their degree of similarity. This Complex graph is stratified into high modularity community partitions by applying the Louvain hierarchical clustering algorithm (Blondel et al., 2008). Due to the non-deterministic nature of the Louvain algorithm, MULTILAYER partitions the graph multiple times (15 events by default) and then selects the most frequent community-partition outcome (the frequency of the community partitions are displayed in the terminal) for their display within the tissue map in which gexels are colored in accordance with their associations with the inferred communities. In addition, the community panel within MULTILAYER displays the list of overexpressed genes comprising the patterns associated with the illustrated communities.

Arguments for Louvain partitioning: (i) Weight: allows inclusion of the similarity index computed within the co-expressed genes as a weight argument, (ii) Multiple iterations: allows the performance of 15 consecutive partitioning operations with the Louvain algorithm and selection of the most highly represented for downstream analyses.

### Gene ontology analysis

MULTILAYER counts using a gene ontology enrichment analysis implemented within the ''gene co-expression pattern detection'' and ''gexel communities'' panels. A collection of GO terms was collected from the Enrichr libraries suite. MULTILAYER infers the confidence of the GO term enrichment by comparing the list of genes from the detected gene co-expression patterns or within a spatial community to those retrieved from the GO database (one-sided Fisher Exact test).

As outcome, MULTILAYER provides a confidence bar plot per enriched GO term, as well as a heatmap matrix displaying the list of genes associated with the enriched GO terms.

### MULTILAYER Compressor ad-hoc module

Multilayer Compressor is an ad-hoc module for generating super-gexel maps by aggregating the raw read counts of contiguous gexels prior to processing. This strategy allows the conversion of a large matrix, such as those retrieved in the case of high-resolution Slide-Seq data (Rodriques et al., 2019), to a compressed format, counting a smaller number of gexels within the grid but with a higher number of read counts per gexel. Multilayer Compressor transforms input data (3 column format composed of gexel coordinate, Gene ID, and read counts per gene), into a data-frame compatible with MULTILAYER (matrix format composed of gexel coordinates in columns and Gene ID in rows) according to the user-defined compression factor parameters (number of gexels on X and Y coordinates). We recommend using the MULTILAYER Compressor when raw ST maps are bigger than 120 x 120 gexel grids.

### Comparison with SPARK and SpatialDE statistical methods

Spatial expression pattern detection readouts for the mouse olfactory bulb and human breast-cancer tissue data (Figure 2) performed with SpatialDE were obtained from https://github.com/xzhoulab/SPARK-Analysis/tree/master/output (Layer2_BC_spe.csv; Rep11_MOB_spe.csv). Furthermore, those generated by SPARK were kindly provided by the laboratory of Dr. Jiaqiang Zhu (see key resources table). For the comparisons, genes were sorted by the adjusted p-value (q-value) descriptor provided by SPDE and compared to those provided by SPARK, as well as with the number of contiguous gexels per pattern detected by MULTILAYER.

### Data availability

All spatial transcriptomics data used within this article were obtained from public repositories and converted into a uniform format compatible with MULTILAYER requirements. Human breast-cancer data (Breast Cancer Layer 2), mouse olfactory bulb data (MOB Replicate 11), published by Stahl et al. (Ståhl et al., 2016), human heart development data, generated by Asp. M. et al. (Asp et al., 2019), and prostate cancer data, generated by Berglund et al. (Berglund et al., 2018), were obtained from the Spatial Research portal. In addition to the raw SrT matrices, their associated hematoxylin and eosin-stained images were also obtained from the same public repository. Hippocampus and cortex high-resolution Slide-seq maps (Rodriques et al., 2019) were obtained from the SpatialDB database (Fan et al., 2020). Pancreatic ductal adenocarcinoma data, generated by Moncada et al. (Moncada et al., 2020), were obtained from the GEO database (GSE111672). Whole mouse embryo DBiT-Seq data, generated by Yang Liu et al (Liu et al., 2020), were obtained from the GEO database (GSE137986). Images corresponding to the original DBiT-Seq article were reproduced with authorization (ELSEVIER License Number 4970301038168).

Compatible MULTILAYER versions of these data are accessible via https://github.com/SysFate/MULTILAYER.

### Code availability

MULTILAYER and MULTILAYER Compressor are available at https://github.com/SysFate/MULTILAYER.